A white paper on nematode comparative genomics

Journal of Nematology 37(4):408–416. 2005.

A White Paper on Nematode Comparative Genomics

David McK. Bird,[1] Mark L. Blaxter,[2] James P. McCarter,[3,4] Makedonka Mitreva,[3] Paul W. Sternberg,[5] W. Kelley Thomas[6]

Abstract: In response to the new opportunities for genome sequencing and comparative genomics, the Society of Nematology (SON) formed a committee to develop a white paper in support of the broad scientific needs associated with this phylum and interests of SON members. Although genome sequencing is expensive, the data generated are unique in biological systems in that genomes have the potential to be complete (every base of the genome can be accounted for), accurate (the data are digital and not subject to stochastic variation), and permanent (once obtained, the genome of a species does not need to be experimentally re-sampled). The availability of complete, accurate, and permanent genome sequences from diverse nematode species will underpin future studies into the biology and evolution of this phylum and the ecological associations (particularly parasitic) nematodes have with other organisms. We anticipate that upwards of 100 nematode genomes will be solved to varying levels of completion in the coming decade and suggest biological and practical considerations to guide the selection of the most informative taxa for sequencing.

Key words: Caenorhabditis elegans, comparative genomics, genome sequencing, systematics.

The "discipline" of genomics arguably began with a project to assemble a physical map of the *Caenorhabditis elegans* genome (Coulson and Sulston, 1984) and certainly was consummated by the attainment of the *C. elegans* genome (The *C. elegans* Sequencing Consortium, 1998). This nematode was the first multicellular organism for which a complete genome sequence was generated, and it remains the only metazoan for which the sequence of every single nucleotide (a total of 100,278,047) has been finished to a high degree of confidence (Chen et al., 2005).

The value of *C. elegans* as a model organism for biomedical research is unquestioned. For example, "the worm" serves as a robust model for complex human traits including Alzheimer's disease (Link et al., 2003), aging (Finch and Ruvkun, 2001; Lee et al., 2003), and diabetes and obesity (McKay et al., 2003). The insight that has been garnered about this species and that is available both in approximately 7,000 publications and via the Web resource, WormBase, elevates *C. elegans* to its status as one of the best understood organisms (Chen et al., 2005; Harris et al., 2004; Stein et al., 2001, 2002; www.wormbase.org).

However, although the genome sequence serves as the glue to integrate much of the collective knowledge on *C. elegans,* it was clear to the *C. elegans* community that the interpretation of its complete sequence would be enhanced by comparison with other related genomes. To this end, a high-quality, 98% complete draft genome sequence for *Caenorhabditis briggsae* was obtained (Stein et al., 2003), providing a platform for comparative genomics. Like *C. elegans,* this species encodes about 20,000 proteins. The coding portions of the *C. elegans* and *C. briggsae* genomes are highly conserved, and essentially all of the known non-coding RNAs are shared between the two species, although non-transcribed regions are highly diverged. Gene order also is conserved, particularly for those genes in operons, but only in local contexts as there have been a large number of mainly within-chromosome rearrangements that break long-range synteny. It is anticipated that identification of conserved elements within this divergent background will point to subtle primary and higher-order regulatory elements across the genome, and this approach has proven effective in gene-by-gene comparisons. However, because genome features evolve at different rates, are of differing sizes, and are detected with varying ease by a variety of tools, comparison of any two species is not sufficient to define many sequence features. To allow better genome alignment, gene interpretation, promoter analysis, and identification of non-coding RNA and other functional features, as well as to explore the forces that mold these genomes, the genomes of three additional *Caenorhabditis* species (*C. remanei, Caenorhabditis* n. sp. PB2801, and *C. japonica*) are currently being obtained. A key to the selection of these taxa is the knowledge of their phylogenetic relationships (Kiontke et al., 2004).

As has been famously noted, "*Caenorhabditis elegans* is a nematode" (Blaxter, 1998), and there is no doubt that this species will serve as an essential guide in exploring the genomes of other nematode species. Conversely, those genomes will aid in the understanding of *C. elegans*. But as detailed below, the interests of the Nematology community at large are broad and varied, and there are many reasons for obtaining additional nema-

Received for publication 19 July 2005.

[1] Center for the Biology of Nematode Parasitism, NC State University, Raleigh, NC 27695.
[2] Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, Edinburgh, EH9 3JT, UK.
[3] Genome Sequencing Center, Department of Genetics, Washington University School of Medicine, St. Louis, MO 63108.
[4] Divergence Inc., St. Louis, MO 63141.
[5] HHMI, Division of Biology, California Institute of Technology, Pasadena, CA 91125.
[6] Hubbard Center for Genome Studies, University of New Hampshire, Durham, NH 03824.

The authors are grateful for discussions with and comments from many colleagues, especially during the Genome Workshop at the 2004 Society of Nematology annual meeting, Estes Park, CO, August 2004 (Bird, 2004), and the Helminth Genome Sequencing Meeting organized by Prof. R. Maizels and held at the Sanger Institute, Cambridge, UK, March, 2004. In particular, we thank Prof. Maizels for allowing us to draw from his report of that meeting in preparing this document.

E-mail: david bird@ncsu.edu

This paper was edited by J. L. Starr.

tode genomes and many reasons for choosing species to be sequenced.

Status of Genome Resources for Nematoda

Molecular phylogenetics (Blaxter et al., 1998; De Ley and Blaxter, 2002) defines three major nematode classes that can be further divided into five clades: Dorylaimia (Clade I), Enoplia (Clade II), Chromadorea and Spirurina (Clade III), Tylenchina (Clade IV), and Rhabditina (Clade V). *Caenorhabditis elegans* is a member of Rhabditina, and the completed and in-progress *Caenorhabditis* genomes (Table 1) anchor this large phylum as the reference Clade V species. The genomes of three additional taxa in Clade V now being sequenced will be particularly informative. One is the strongylid animal parasite *Haemonchus contortus*. The second is *Pristionchus pacificus,* a free-living nematode that has proven to be a powerful laboratory model for comparative development (Eizinger and Sommer, 1997) and may be particularly informative because of its relatively basal position in Clade V. The third is the entopathenogenic species *Heterorhabditis bacteriophora,* which not only will shed light on bacterial-nematode and insect-nematode interactions (the symbiotic bacterial partner of *H. bacteriophora* has also been sequenced) but, because it has been proposed that *Caenorhabditis* themselves have insect associations (Blaxter and Bird, 1997), may give clues to *C. elegans* ecology.

A draft genome sequence of *Brugia malayi,* which is responsible for human filariasis and elephantiasis, has recently been obtained (Ghedin et al., 2004) and is the first parasitic nematode to be sequenced (Table 1). Like *C. elegans, B. malayi* has six chromosomes, which comprise the estimated 90 Mb genome. Using a whole genome shotgun strategy in conjunction with sequencing the ends of large-insert, Bacterial Artificial Chromosome (BAC) clones, the project has generated 9-fold genome coverage. Although the annotation phase of this genome has only just begun, previous analyses of *Brugia* ESTs (Bird et al., 1999) and longer segments of the *Brugia* genome (Guiliano et al., 2002) have revealed the utility of *C. elegans* as an annotation platform. Not surprisingly, given that these nematodes last shared a common ancestor more than 300 million years ago, the degree of synteny is not extensive across the genome (Ghedin et al., 2004) but exhibits some local conservation (Guiliano et al., 2002); intra-chromosomal rearrangement is greatly favored over inter-chromosomal rearrangement (Whitton et al., 2004). In addition to contributing to an enhanced understanding of filarial parasite biology and suggesting new vaccine candidates and drug targets, *Brugia* defines a reference clade III nematode genome (Blaxter et al., 1998) and, like the *Caenorhabditis* genomes, will be invaluable for annotating additional nematode genomes (Bird and Opperman, 1998; Bird et al., 1999). Clade III comprises multiple orders of animal parasitic taxa, including Spirurida, Oxyurida, Ascarida, and Rhigonematida. A key to understanding the radiation of this entirely parasitic clade will be comparative genomic sequence from the most appropriate free-living outgroup, such as *Plectus acuminatus*.

The U.S. Department of Agriculture (USDA) has recently committed to funding a deep draft genome sequence of the root-knot nematode, *Meloidogyne hapla,* as the first plant-parasitic nematode species to be sequenced (Table 1). *Meloidogyne hapla* will serve as the representative Clade IV nematode genome and, just as the *Brugia* and *Caenorhabditis* sequences collectively will aid annotation of the *Meloidogyne* sequence, in turn will help the annotation of those, thus becoming a model unto itself. Funding also has been secured for the vertebrate parasites *Trichinella spiralis* (clade I) and *Haemonchus contortus* (clade V) (Table 1).

In addition to genome sequencing, many nematode genes have been identified from expressed sequence tag (EST) projects (McCarter et al., 2003a). ESTs are

Table 1. Genome sizes and chromosome numbers of nematode taxa for which a genome project is under way.

Species	Clade[1]	Tropic ecology[2]	Type of genome project	Status[3]	Genome size (Mb)	Funding[4]
Caenorhabditis elegans	V	B	Full genome sequence	C	100.3	NHGRI
Caenorhabditis briggsae	V	B	Whole genome draft	C	105	NHGRI
Caenorhabditis remanei	V	B	Whole genome draft	C	~140	NHGRI
Caenorhabditis japonica	V	B	Whole genome draft	P	—	NHGRI
Caenorhabditis sp. c.f. PB2801	V	B	Whole genome draft	P	—	NHGRI
Pristionchus pacificus	V	A-O-P	Whole genome draft	P	~110	NHGRI
Heterorhabditis bacteriophora	V	I	Whole genome draft	P	~110	NHGRI
Brugia malayi	III	V	Whole genome shotgun	C	~100	NIAID
Haemonchus contortus	V	V	Whole genome draft	P	~55	Sanger
Meloidogyne hapla	IV	P	Pooled BAC sequencing	P	50	CSREES
Trichinella spiralis	I	V	Whole genome draft	P	270	NHGRI

[1] Clade based on the assignment of Blaxter et al., 1998.
[2] Food source. B: Bacteriovore; A-O-P: Algivore-Omnivore-Predator; I: Insect-associated bacteriovore; V: Vertebrate parasite; P: Plant parasite.
[3] Status of genome sequencing project. C: Completed; P: Planned or in progress.
[4] Funding source for genome sequencing project. NHGRI: National Human Genome Research Institute, United States; NIAID: National Institute of Allergy and Infectious Disease, United States; Sanger: The Welcome Trust Sanger Institute, United Kingdom; CSREES: NSF/USDA CSREES Microbial Genome Sequencing Program, United States.

single-pass sequencing scans of randomly selected cDNA clones (McCarter et al., 2000). As recently as 2000 there were only 24,000 ESTs in public databases from nematodes other than *C. elegans,* but by December 2004 nearly 350,000 had been deposited, disproportionately focusing on parasites of humans (e.g., Blaxter et al., 2002; Daub et al., 2000), animals (e.g., Tetteh et al., 1999), and plants (e.g., Bird et al., 2002; McCarter et al., 2003b; Mitreva et al., 2004). A meta-analysis of the genomic biology of the phylum Nematoda was completed using >250,000 ESTs originating from 30 species, clustered into 93,000 genes and grouped into 60,000 gene families (Parkinson et al., 2004). This collection of data was used to estimate the degree to which "genespace" (the diversity of distinct genes) within nematodes has been sampled. In nematodes, despite the availability of the genomes of two *Caenorhabditis* species, genespace appears far from thoroughly sampled, as the addition of new species to the analysis has yielded a linear increase in discovery of new genes. Therefore, despite a deceptively uniform body plan, nematodes seem to be more diverse at the molecular level than was previously recognized. The set of ~20,000 genes and 12,000 gene families represented by *C. elegans* provides a starting point for exploring this diversity and does capture many of the conserved gene families shared with other eukaryotes, but it represents only a small portion of the expanding total nematode genespace. As sequencing has been performed from only a few dozen of perhaps over one million nematode species, representing a very limited component of the phylogenetic and ecological diversity of the phylum, the vast majority of nematode genespace remains unsampled. Free-living taxa representing basal lineages in the phylum (Clades I and II) are particularly critical in this respect.

Goal of This Paper

As a group, we recognize two important facts. First, Nematology is a discipline as broad as its phylum and spans numerous and diverse fundamental scientific goals, including exploiting the *C. elegans* model for basic and biomedical biology, agricultural interests associated with plant- and animal-parasitic groups, medical and veterinary interests in species that impact human and animal health, and exploring the vast biodiversity and ecological associations of the Nematoda. Second, sequencing of genomes is now a basic approach to addressing questions in biology and will continue to expand in practice. Numerous research approaches from gene expression profiling (microarrays) to proteomics are built upon an assumption of an available genome sequence and reasonably accurate gene models. Consequently, we need to prepare for a future where large numbers of nematode genomes will be sequenced. Considering these two facts, it seems that the immediate need is to organize an approach that will foster the development of many new nematode genome sequences from across the phylum and disciplines. Ultimately, the challenge is to promote the sequencing of the most informative set of nematode genomes that supply the tools for functional genomics and allow for informative comparative genomics across the phylum.

It is important to stress that, other than for the sake of example and to make broad generalities, we purposely avoid prioritizing nematode taxa. This is not for political expediency but because there is no need. Prioritization is specific to the user community and relevant funding agencies and is not a useful phylum-wide exercise. Rather, we outline fundamental needs within Nematology and the justification in each context for

Box 1. Criteria for selecting nematode species for sequencing.

1. Importance of the nematode
   a. Parasite
      i. Human
         1. Major parasites of interest for drug, vaccine, or diagnostic development
         2. Parasites of interest as immune/hematopoietic modulators
         3. Models of human parasites (e.g., parasites of rodents)
      ii. Animal
         1. Farm animals
            a. Livestock
            b. Poultry
            c. Fish
         2. Companion animals
         3. Invertebrates
      iii. Plant
      iv. Virus or bacterial vector
   b. Tractable model for a parasite
      i. Ease of culture
      ii. Manageable host
      iii. Availability of forward genetics
      iv. Ability to transform or use RNAi
      v. Availability/accessibility of developmental/larval stages
   c. "Useful" nematodes
      i. Biocontrol agents
      ii. Insecticidal species
      iii. Saprophytes
   d. Phylogenetic position
      i. Basal species for the entire phylum
      ii. Representative of under-represented clade
      iii. Species to help understand relationships of parasitic species (e.g., free-living relative)
   e. Ecologically significant species
      i. Environmental monitors
      ii. Representatives of particular niches
         1. Marine nematodes
         2. Extremophiles
   f. Annotation models
      i. Relationship to *C. elegans*
      ii. Relationship to important parasites
2. Beneficiaries of the genome sequence
   a. A significant, existing community of researchers able to exploit the genome information
   b. The potential for whole genome sequence to stimulate research activity in a neglected area (e.g., a disease-causing pathogen) and (or) the potential to facilitate new entrants commencing work on nematodes
3. Mechanism for using the sequence
   a. Annotation plans
   b. Dissemination of resources
   c. Computational power

additional nematode genome sequences. We recommend the development of consortia and focused user communities to support specific proposals; each species needs a champion. We broadly list some criteria that such communities may consider in selecting taxa for genome sequencing (Box 1), and we discuss some of the practical considerations at length. Finally, we propose the development of a database of nematode species "*vitae*" to document features of those species that support their use in comparative genomics.

Broad Topics to be Addressed by Nematode Genome Sequencing

Basic biological processes that define, distinguish, and differentiate the Nematoda: The advent of genomic analyses of model organisms, including whole genome sequencing and extensive profiling of transcript patterns using microarrays, has begun to reveal that key biological processes are strikingly more widely conserved across life than has previously been imagined. It has, for example, proven possible to identify large sets of genes broadly conserved across the Eukaryota (including yeast, *C. elegans, Arabidopsis, Drosophila,* and human), leading to inference of a presumptive ancestral eukaryotic genome (Koonin et al., 2004). Even more remarkable than finding gene sets that are conserved across kingdoms has been the realization that patterns of gene expression are conserved. McCarroll et al. (2004) have developed tools to use the expression profile of a panel of genes that show coordinate behavior during a particular process (such as aging) in one species as a query to sets of gene expression data from other species. As an example of the power of this approach, using the profiles of yeast genes with altered expression during sporulation to query the collection of *C. elegans* gene profiles gave the strongest match to nematode genes involved in germ line proliferation (McCarroll et al., 2004). In other words, the pattern of gene regulation controlling meiosis/mitosis in a single-celled eukaryote is detectably and specifically similar to that in a more complex, multicellular organism.

However, by their nature, the wide, cross-phylum comparisons of genes undertaken to date have focused on genes that are broadly conserved and likely to encode core eukaryotic functions. Although a role in the specific biology unique to the particular species cannot be ruled out for such core genes, it seems more likely that other, more divergent genes serve to make particular species unique. A comparison between *C. briggsae* and *C. elegans* may provide some clues. Although almost indistinguishable by light microscopy and apparently sharing identical biology as assayed in the laboratory, these species split from a common ancestor about 100 million years ago (see Kiontke et al., 2004, for discussion of divergence rates). Approximately 12,200 *C. briggsae* genes can be assigned unequivocally as *C. elegans* orthologs and are presumed to encode the same functions. A further 6,500 have one or more clearly detectable *C. elegans* homologs (i.e., have arisen from a common ancestral gene) but may have diverged following gene duplication (i.e., be paralogs) and adopted different functions. However, even employing a fairly non-stringent standard for what constitutes a match (BLAST $\leq 1.0 \times e^{-5}$), 807 *C. briggsae* genes (4.1%) have no detectable match in *C. elegans*. Conversely, 5.1% of the *C. elegans* gene set failed to match a *C. briggsae* gene. Understanding these genes in the context of each species' biology will likely prove interesting.

Given what we know about gene content differences between *C. briggsae* and *C. elegans* and the emerging picture from EST analysis that the genespace across the Nematoda is surprisingly large (Mitreva et al., 2005a; Parkinson et al., 2004), it will be especially important to understand what sorts of new genes nematodes are evolving, especially for parasites where specific and probably unique selection pressures are present. The broad conservation of gene content across the eukaryotes suggests that it might prove productive to change research emphasis from attempting to understand what is similar between two species to attempting to identify the particular molecular genetic distinctions responsible for making the organisms different. This goal places an emphasis on the generation of complete genome sequences to distinguish between unique and shared gene sets among taxa. Partial genome sequences can provide insights into shared sets of genes but cannot confirm that a gene (set) is missing without the complete sequence. Furthermore, within the context of such a diverse phylum, it will be critical to select taxa at appropriate levels of divergence and within a well-defined phylogenetic context to distinguish between loss of a gene, gain of a gene, and rapid divergence of a gene.

Comparative genomics to understand parasitism: As noted, genome-wide comparisons will prove very useful for developing and understanding the genetic differences and similarities correlated with specific biological attributes including evolution of parasitism. The phylogeny (Blaxter et al., 1998) supports multiple origins of parasitism in Nematoda, and it is reasonable to expect that different strategies and molecular innovations may underlie adaptations in different lineages. Mechanisms that could affect evolution to parasitism include gene duplication and diversification, gene-loss, changes in patterns of gene expression (Denver et al., 2005), alterations in genes controlling metabolic and developmental functions, adaptation of pre-existing genes to encode new functions, and acquisition of genes from other species via horizontal gene transfer (HGT). Accumulating evidence supports a bacterial origin for some genes in Tylenchid plant-parasitic nematodes. Among those proposed are genes encoding enzymes that can degrade two major components of plant cell

walls, as well as genes with potential roles in host-parasite signaling (Davis and Mitchum, 2005). Most of these genes were identified on the basis of biochemical or immunological criteria, with claims of HGT being supported by phylogenetic incongruence. A bioinformatic approach using phylogenetic filters identified those genes previously found, plus an additional set of candidate HGT genes (Scholl et al., 2003). However, one problem with these approaches is that, in the absence of complete genome sequences, the (EST) datasets are incomplete; the absence of evidence of a particular gene in such a set is not truly evidence for absence. Completed genomes will serve both as a source of information on the presence and absence of genes and for the development of functional genomic and proteomic tools to foster rapid and cost-effective biochemical studies. The utility of such approaches can be seen in the discovery of novel anthelminthics (McCarter, 2004).

Nematodes parasitic on vertebrates also have adapted for interaction with their hosts, including evolving means of evading or modulating the host immune and hematopoietic responses. The study of parasite-encoded proteins that interact with the mammalian immune system has provided numerous insights of importance to the broader field of immunology (Maizels and Yazdanbakhsh, 2003). Parasitic nematodes are currently being tested in clinical trials as direct therapeutics for autoimmune diseases (e.g., *Trichuris suis* for inflammatory bowel disease) (Hunter and McKay, 2004; Summers et al., 2005). Similarly, a hookworm-derived recombinant protein that inhibits human FVIIa/TF is in clinical trials, in this case as an anticoagulant for unstable angina and myocardial infarction (Lee and Vlasuk, 2003; Mungall, 2004; Stanssens et al., 1996). Further understanding of parasite molecular mediators of host interactions has promise to lead to additional human therapeutics with applications well beyond parasitology and nematology.

Comparative genomics and the need for a well-supported evolutionary framework: Sampling from a relatively small number of nematode taxa broadly spanning the phylum gave a highly informative view of the phylogenetic structure of Nematoda (Blaxter et al., 1998). Whole genome information from broadly selected nematode taxa will be needed to resolve the deepest branches in the phylum. Such efforts are concordant with the Nematode Tree-of-Life Project (NemATOL: http://nematol.unh.edu/) and, conversely, each nematode to be sequenced will have to be placed in a phylogenetic context for effective use of its data. In addition to better understanding evolutionary relationships within the phylum, nematode genome data will contribute to understanding the relationship of Nematode to other metazoan phyla, a point that remains controversial (Blair et al., 2002). Recent data (Philippe et al., 2005) are consistent with the model that places nematodes together with arthropods into the clade Ecdysozoa (Aguinaldo et al., 1997). Part of the reason for the uncertainty in establishing the true position of nematodes within the animal kingdom stems from long-branch attraction artifacts (Felsenstein, 1978) that are exacerbated by high evolutionary rates for genes used in phylogenetic inference (Philippe et al., 2005). One solution to this problem is to dissect the branches by including multiple taxa for each clade (Hendy and Penny, 1989). Addressing the fundamental question of where nematodes fit in animal evolution will greatly benefit from the inclusion of multiple, diverse nematode genome sequences.

Understanding the evolutionary relationships between taxa is essential for comparative annotation. One of the most powerful ways of understanding the "meaning" of a DNA sequence is to look for the consequences of natural selection after evolutionary divergence. By sequencing multiple taxa that are sufficiently different, the conservation of sequence becomes informative with respect to function. As noted above, this is the motivating logic for the various *Caenorhabditis* projects and should serve as a model for all nematode sequencing projects. As more genomes are analyzed, it has become possible to predict just how related genomes need to be in order to identify functionally conserved domains by comparison (Eddy, 2005).

Comparative genomics to support ecological and evolutionary functional genomics: Ecological and evolutionary functional genomics is an emerging field intent upon developing systems to bring functional genomics to ecological studies. As such, it is expected that one key motivation for the sequencing of nematode genomes will be to support the development of tools for functional genomics. DNA sequences are critical in the development of microarray platforms for the analysis of gene expression, for the interpretation of proteomic studies, and in the discovery of polymorphisms for the analysis of quantitative trait loci. Because it is unlikely that many nematode species will be directly examined by traditional biological means, genomic and post-genomic tools suggest an approach to consider nematodes and their ecological associations as a group (a process termed "metagenomics").

Practical Criteria for Selecting Nematode Species for Sequencing

Perhaps the most important point to address is "who will use the sequence?" Fortunately, the breadth of interests of the nematology community suggests that each sequence might be of interest to multiple constituencies. For example, the sequence of a human-parasitic species would obviously be of interest to those labs working on the nematode species in question and likely also to clinicians interested in the pathology. But further, each new species provides a phylogenetically in-

formative platform and may contribute additional functional information relevant to the annotation of *C. elegans* and the rest of the phylum. For example, the discovery of root-knot nematode ESTs with highly significant matches to hypothetical genes predicted by automated annotation of the *C. elegans* genome strongly implies that those predictions indeed define genes, albeit with no known function (McCarter et al., 2003b). Hookworm ESTs have identified orthologs of genes that were previously "orphans" in either *C. elegans* or *C. briggsae* (Mitreva et al., 2005b). It is possible that future meta-analyses of nematode genomes will reveal classes of genes associated with particular biological attributes (such as parasitism or, more generally, symbiosis) (Ott et al., 2004a, 2004b). The incorporation of non-*C. elegans* nematode sequences into an extensive information management system such as WormBase or a system for displaying phylogenetic context (e.g., NemATOL) is an important step toward a more effective exploitation of comparative nematode data.

It is important to consider the specific scientific needs and size of the research community associated with any proposed nematode genome sequencing project. Downstream of a large-scale sequencing project are the processes of assembly and annotation. Some sequencing organizations can dedicate post-docs or experienced annotators for this activity, but the issue must be addressed in any overall whole-genome sequencing proposal. Much of the finishing and annotation can be automated in the first instance, but it is likely that individual research communities will be responsible for subsequent manual annotation. It also is important to identify post-genomic activities. Functional genomic resources such as microarrays, cDNA archives, and libraries offer a post-genomic pipeline to leverage the sequences and are best organized and most cost-effective as large-scale cooperative efforts by the user community. The availability of tools, such as RNA interference (Fire et al., 1998), for downstream analysis of gene function might further influence species selection.

Genome readiness: It is perhaps self-evident that sufficient, high-quality genomic DNA, free from contamination by other species (except for purposeful metagenomic analysis) be available for library construction. The technical constraints of constructing large insert libraries (>100 kb inserts) generally dictate that sufficient intact nuclei able to yield hundreds of micrograms of DNA be available. For some species that might otherwise be assigned high priority for sequencing (such as certain animal parasites, or phylogenetically significant species), the fact that they live in difficult-to-sample habitats should unfortunately eliminate them from consideration, given current technology, unless the community dedicates significant effort to obtain materials for sequencing. For example, it remains true that the vast majority of nematodes, in particular those representing basal lineages and under-sampled clades in the phylum, are not available as laboratory cultures. The establishment of such cultures would be an important step to overcome these practical issues. Current efforts focusing on the culture of *Tobrilus* species (De Ley, pers. comm.) and published accounts of the possibility of similar cultures (Moens and Vincx, 1998) suggest that some of these practical limitations can be overcome.

It also is true that genome size matters, as sequencing costs are directly proportional to the number of bases that must be obtained. Genome size must first be assessed, ideally using independent methods such as Feulgen image analysis densitometry (Hardie et al., 2002) or flow cytometry (Kent et al., 1998). Other parameters, such as G + C content and complexity (i.e., the proportion of repeats in the genome), are important in evaluating the potential success of a proposed sequencing project. For assembling sequences obtained by the whole genome shotgun (WGS) approach, which is now a favored method for draft-quality genomes, an important criterion is the level of polymorphism in a particular genome, which is reflected in the heterogeneity of the genome sample in a particular isolate. A high degree of heterozygosity may hamper assembly of WGS sequence; so, unless the level of polymorphism is naturally low, highly inbred strains are desirable. The most recent advances in genome sequencing technology that involve highly parallel sequencing of hundreds of thousands of templates simultaneously are likely to dramatically reduce the cost of sequencing. But issues such as purity of sample and levels of polymorphism will remain critical criteria in taxa selection.

Supporting tools: No matter what sequencing strategy is followed (WGS or a more directed approach), the availability of anchored physical and genetic maps can provide a framework to assist with genome assembly. Similarly, the availability of cDNA libraries for EST coverage is necessary for gene identification and prediction of exon-intron boundaries. Ideally, ESTs from each stage of complex life cycles should be included. Some full-length cDNAs are required for training gene prediction programs such as GLIMMER (Salzberg et al., 1998).

Nematode Vitae

We suggest that, as a standardized format to capture the types of genome information, communities of nematode researchers establish "vitae" for nematodes to be sequenced. Ideally, such vitae should be freely available in a database (at a Web site, for example). For each species, a general overview is provided, followed by the significance or reason why the species should be sequenced. A general description (e.g., of the biology, pathology, or ecology) is provided as well as the known "genome facts," such as the size of the genome, available EST or map resources, availability of libraries, etc.

An indication of the size of the community interested in the nematode species is given, with a supporting argument (e.g., the number of publications). Other details, including how the genome sequence will be used, are presented as are contact details, such as Web sites, consortia leaders, etc. It is intended that the vitae encapsulate the type of information necessary to justify to a funding agency the need to sequence the particular genome. An example, using the animal-parasitic nematode *Haemonchus contortus,* is shown.

Haemonchus contortus Vita

Overview: Haemonchocus contortus is the most economically important nematode parasite of sheep and goat worldwide. Control of this parasite is increasingly difficult due to widespread resistance to all the major classes of anthelmintic drugs throughout the world. This parasite has the largest active research community of all clade V nematodes and arguably has been the subject of more research on vaccine development and drug resistance than any other parasitic nematode species. Hence it can be considered to be a model parasitic nematode as well as an important pathogen in its own right. Importantly, it is related to the devastating human hookworm pathogens (*Necator americanus* and *Ancylostoma duodenale*) and will thus illuminate their biology and promote control options for human disease. It is a relatively large adult parasite (2 cm) with high fecundity, permitting the relatively straightforward generation of large quantities of parasite material for biochemical, immunological, and molecular studies. It is possible to perform genetic crosses between isolates, and already there is a genome mapping project funded by the Wellcome Trust to produce an integrated HAPPY and BAC clone map and develop polymorphic markers for genetic analysis. HAPPY mapping (Dear and Cook, 1993) is a physical mapping approach that generates information that is analogous to genetic linkage data. The integrated HAPPY/BAC clone map of *H. contortus* will be of significant value for contig assembly of a full genome sequencing project. There are also 21,967 ESTs available, representing an estimated 4,000 genes, which will be invaluable for gene finding and annotation. RNAi is being developed in this species at a number of sites worldwide, and success has been reported with a number of different gene targets.

Significance: Many clade V nematodes have veterinary impact. Of most importance are the parasites of grazing livestock that cause significant economic problems for agriculture and detrimental effects on animal welfare throughout the world. A recent report, commissioned by the United Kingdom Department for International Development, listed the trichostrongylid nematodes (ranked as a group because they often occur as mixed infections) at the top of a list of the top 80 animal diseases that have a major impact on the poor in the developing world, i.e., they are considered to have a greater impact than any other disease of domestic animals. This group of parasites is among the most economically important diseases of livestock in the developed world.

General description: Many different species, from different genera, occur as mixed infections of livestock. The relative importance of particular species varies for different regions of the world; consequently, the relative priority of each will differ between funding agencies. The disease syndrome observed varies depending on the species predominating. For example, the parasite *H. contortus* is a blood-feeder and, hence, highly pathogenic. Others cause mild clinical disease, but in all cases subclinical infections can dramatically reduce productivity in livestock units. Sheep nematodes provide an important resource in that they are readily passaged and relatively inexpensive to maintain; some of the greatest success in nematode vaccine development has occurred in this sector. In addition, it is a priority to elucidate the genetic basis of anthelmintic resistance in these species. We propose that *H. contortus* represents the most appropriate species from this group for full sequencing and assembly (10X coverage, finishing, assembly, and annotation). Six other trichostrongylid nematode species of veterinary importance are suggested for 5X coverage. These species have been chosen as the most economically important and tractable species from each of the remaining genera of primary importance.

Genome facts: The genome size of *Haemonchus* is estimated to be 54Mb, and all the indications are that the genome sizes of this entire group are in this region. RNA interference is effective in L3 *Haemonchus,* and BAC library development and HAPPY mapping are ongoing. As noted for the individual species below, varying EST datasets are available, including from *Haemonchus* (17,269 ESTs from approximately 4,145 genes expressed in adult, L3, and L4 worms), *Teladorsagia* (4,379 ESTs from approximately 1,700 genes), and *Ostertagia ostertagi* (7,600 ESTs from 2,350 genes expressed in larval stages.)

Community and active labs: Many labs worldwide study gastrointestinal nematodes of livestock, reflecting their global importance. A key word search of PubMed revealed 1,800 references with the query term "Haemonchus," 980 references with the query "Ostertagia/Teladorsagia," 1,448 matches to the query "Trichostrongylus" and 796 and 639 to "Dictyocaulus" and "Cooperia," respectively.

Curation: The *Haemonchus* and *Teladorsagia* ESTs are available individually and grouped into clusters at Nembase (www.nematodes.org), along with additional annotation. The individual clones are freely available to the community, and initial publications describing the datasets are available (Geldhof et al., 2005). *Haemon-*

chus contortus has a project page on the Sanger Centre at http://www.sanger.ac.uk/Projects/H_contortus/.

Exploitation: The genome information generated will have immediate and urgent application in the identification of novel target molecules for the control of these parasites by vaccination or by drug development. In addition, these datasets will be invaluable for characterizing the mechanisms involved in drug resistance. The community has available selected lines of some species with defined resistance to all the currently available anthelmintics. This will prove an invaluable resource for comparative genomics to seek common genes involved with this problem. Having defined genome sequences opens the door for comprehensive and meaningful analyses of gene expression, e.g., by microarrays.

Benefits: The benefits would be to accelerate the development of novel methods to control these parasites in livestock, defining the genetic mechanisms underlying drug susceptibility and resistance with the possibility of extending the useful life of existing drugs and improved diagnostics in this area, as well as providing the means for meaningful whole-animal studies of the host-parasite interaction.

Conclusion

As the jigsaw that was the physical map of the *C. elegans* genome began to take shape with more than 40% of the genome in multiple contigs, Coulson and Sulston (1984) observed: "We've done the straight edges and the little house in the middle and we're working on the sky. There's quite a lot of it." As more nematode genomes are solved, each will serve to produce a clearer picture on the lid of the box to guide the genomes to follow. All indications are that the "sky" will reveal a diverse and large genespace spanning the Nematoda, which will underpin the acceleration of hypothesis-driven research stemming from nematode genome informatics.

Literature Cited

Aguinaldo, A. M., J. M. Turbeville, L. S. Linford, M. C. Rivera, J. R. Garey, R. A. Raff, and J. A. Lake. 1997. Evidence for a clade of nematodes, arthropods, and other moulting animals. Nature 387:489–493.

Bird, D. McK. 2004. High society (of nematologists). Genome Biology 5:353.

Bird, D. McK., S. W. Clifton, T. Kepler, J. J. Kieber, J. Thorne, and C. H. Opperman. 2002. Genomic dissection of a nematode-plant interaction: A tool to study plant biology. Plant Physiology 129:394–395.

Bird, D. McK., and C. H. Opperman. 1998. *Caenorhabditis elegans:* A genetic guide to parasitic nematode biology. Journal of Nematology 30:299–308.

Bird, D. McK., C. H. Opperman, S. J. M. Jones, and D. L. Baillie. 1999. The *Caenorhabditis elegans* genome: A guide in the post genomics age. Annual Review of Phytopathology 37:247–265.

Blair, J. E., K. Ikeo, T. Gojobori, and S. B. Hedges. 2002. The evolutionary position of nematodes. BMC Evolutionary Biology 2:7.

Blaxter, M. L. 1998. *Caenorhabditis elegans* is a nematode. Science 282:2041–2046.

Blaxter, M. L., and D. McK. Bird. 1997. Parasitic nematodes. Pp. 851–878 *in* D. L. Riddle, T. Blumenthal, B. Meyer, and J. Priess, eds. *C. elegans* II. Cold Spring Harbor, NY: Cold Spring Harbor Press.

Blaxter, M., J. Daub, D. Guiliano, J. Parkinson, C. Whitton, and The Filarial Genome Project. 2002. The *Brugia malayi* genome project: Expressed sequence tags and gene discovery. Transactions of the Royal Society of Tropical Medicine and Hygiene 96:7–17.

Blaxter, M. L., P. DeLey, J. Garey, L. X. Liu, P. Scheldeman, A. Vierstraete, J. R. Vanfleteren, L. Y. Mackey, M. Dorris, L. M. Frisse, J. T. Vida, and W. K. Thomas. 1998. A molecular evolutionary framework for the phylum Nematoda. Nature 392:71–75.

Chen, N., T. W. Harris, I. Antoshechkin, C. Bastiani, T. Bieri, D. Blasiar, K. Bradnam, P. Canaran, J. Chan, C. K. Chen, W. J. Chen, F. Cunningham, P. Davis, E. Kenny, R. Kishore, D. Lawson, R. Lee, H. M. Muller, C. Nakamura, S. Pai, P. Ozersky, A. Petcherski, A. Rogers, A. Sabo, E. M. Schwarz, K. Van Auken, Q. Wang, R. Durbin, J. Spieth, P. W. Sternberg, and L. D. Stein. 2005. WormBase: A comprehensive data resource for *Caenorhabditis* biology and genomics. Nucleic Acids Research 33:D383–389.

Coulson, A., and J. Sulston. 1984. The genomic jigsaw. Worm Breeder's Gazette 8(3):6.

Davis, E. L., and M. G. Mitchum. 2005. Nematodes. Sophisticated parasites of legumes. Plant Physiology 137:1182–1188.

Daub, J., A. Loukas, D. I. Pritchard, and M. Blaxter. 2000. A survey of genes expressed in adults of the human hookworm, *Necator americanus.* Parasitology 120:171–184.

Dear, P. H., and P. R. Cook. 1993. HAPPY mapping: Linkage mapping using a physical analogue of meiosis. Nucleic Acid Research 21:13–20.

De Ley, P., and M. L. Blaxter. 2002. Systematic position and phylogeny. Pp. 1–30 *in* D. Lee, ed. The biology of nematodes. London: Taylor and Francis.

Denver, D. R., K. Morris, J. T. Streelman, S. K. Kim, M. Lynch, and W. K. Thomas. 2005. The transcriptional consequences of mutation and natural selection in *Caenorhabditis elegans.* Nature Genetics 37: 544–548.

Eddy, S. R. 2005. A model of the statistical power of comparative genome sequence analysis. PLoS Biology 3:e10.

Eizinger, A., and R. J. Sommer. 1997. The homeotic gene *lin-39* and the evolution of nematode epidermal cell fates. Science 278:452–455.

Felsenstein, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Systematic Zoology 27:401–410.

Finch, C. E., and G. Ruvkun. 2001. The genetics of aging. Annual Review of Genomics and Human Genetics 2:435–462.

Fire, A., S. Xu, M. K. Montgomery, S. A. Kostas, S. E. Driver, and C. C. Mello. 1998. Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans.* Nature 391:806–811.

Geldhof, P., C. Whitton, W. F. Gregory, M. Blaxter, and D. P. Knox. 2005. Characterisation of the two most abundant genes in the *Haemonchus contortus* expressed sequence tag dataset. International Journal for Parasitology 35:513–522.

Ghedin, E., S. Wang, J. M. Foster, and B. E. Slatko. 2004. First sequenced genome of a parasitic nematode. Trends in Parasitology 20:151–153.

Guiliano, D. B., N. Hall, S. J. Jones, L. N. Clark, C. H. Corton, B. G. Barrell, and M. L. Blaxter. 2002. Conservation of long-range synteny and microsynteny between the genomes of two distantly related nematodes. Genome Biology 3:R57.

Hardie, D. C., T. R. Gregory, and P. D. N. Hebert. 2002. From pixels to picograms: A beginner's guide to genome quantification by Feulgen image analysis densitometry. Journal of Histochemistry and Cytochemistry 50:735–749.

Harris, T. W., N. Chen, F. Cunningham, M. Tello-Ruiz, I. Antoshechkin, C. Bastiani, T. Bieri, D. Blasiar, K. Bradnam, J. Chan, C. K. Chen, W. J. Chen, P. Davis, E. Kenny, R. Kishore, D. Lawson, R. Lee, H. M. Muller, C. Nakamura, P. Ozersky, A. Petcherski, A. Rogers, A. Sabo, E. M. Schwarz, K. Van Auken, Q. Wang, R. Durbin, J. Spieth, P. W. Stenberg, and L. D. Stein. 2004. WormBase: A multi-species resource for nematode biology and genomics. Nucleic Acids Research 32:D411–417.

Hendy, M., and D. Penny. 1989. A framework for the quantitative study of evolutionary trees. Systematic Zoology 38:297–309.

Hunter, M. M., and D. M. McKay. 2004. Review article: Helminths as therapeutic agents for inflammatory bowel disease. Alimentary Pharmacology & Therapeutics 19:167–177.

Kent, M., R. Chandler, and S. Wachtel. 1988. DNA analysis by flow cytometry. Cytogenetics and Cell Genetics 47:88–89.

Kiontke, K., N. P. Gavin, Y. Raynes, C. Roehrig, F. Piano, and D. H. Fitch. 2004. *Caenorhabditis* phylogeny predicts convergence of hermaphroditism and extensive intron loss. Proceedings of the National Academy of Sciences 101:9003–9008.

Koonin, E. V., N. D. Fedorova, J. D. Jackson, A. R. Jacobs, D. M. Krylov, K. S. Makarova, R. Mazumder, S. L. Mekhedov, A. N. Nikolskaya, B. S. Rao, I. B. Rogozin, S. Smirnov, A. V. Sorokin, A. V. Sverdlov, S. Vasudevan, Y. I. Wolf, J. J. Yin, and D. A. Natale. 2004. A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. Genome Biology 5:R7.1–R7.27.

Lee, A. Y. Y., and G. P. Vlasuk. 2003. Recombinant nematode anticoagulant protein c2 and other inhibitors targeting blood coagulation factor VIIa/tissue factor. Journal of Internal Medicine 254:313–321.

Lee, S. S., S. Kennedy, A. C. Tolonen, and G. Ruvkun. 2003. DAF-16 target genes that control *C. elegans* life-span and metabolism. Science 300:644–647.

Link, C. D., A. Taft, V. Kapulkin, K. Duke, S. Kim, Q. Fei, D. E. Wood, and B. G. Sahagan. 2003. Gene expression analysis in a transgenic *Caenorhabditis elegans* Alzheimer's disease model. Neurobiology of Aging 24:397–413.

Maizels, R. M., and M. Yazdanbakhsh. 2003. Immune regulation by helminth parasites: Cellular and molecular mechanisms. Nature Reviews: Immunology 3:733–744.

McCarroll, S. A., C. T. Murphy, S. Zou, S. D. Pletcher, C. S. Chin, Y. N. Jan, C. Kenyon, C. I. Bargmann, and H. Li. 2004. Comparing genomic expression patterns across species identifies shared transcriptional profile in aging. Nature Genetics 36:197–204.

McCarter, J. P. 2004. Genomic filtering as an approach to discovering novel antiparasitics. Trends in Parasitology 20:462–468.

McCarter, J. P., P. Abad, J. Jones, and D. McK. Bird. 2000. Rapid gene discovery in plant-parasitic nematodes via Expressed Sequence Tags. Nematology 2:719–731.

McCarter, J. P., M. Mitreva, S. W. Clifton, D. McK. Bird, and R. Waterston. 2003a. Nematode gene sequences: Update for December 2003. Journal of Nematology 35:465–469.

McCarter, J. P., M. D. Mitreva, J. Martin, M. Dante, T. Wylie, U. Rao, D. Pape, Y. Bowers, B. Theising, C. Murphy, A. P. Kloek, B. Chiapelli, S. W. Clifton, D. McK. Bird, and R. Waterston. 2003b. Analysis and functional classification of transcripts from the root-knot nematode *Meloidogyne incognita*. Genome Biology 4:R26.1–R26.19.

McKay, R. M., J. P. McKay, L. Avery, and J. M. Graff. 2003. *C. elegans:* A model for exploring the genetics of fat storage. Developmental Cell 4:131–142.

Mitreva, M., M. L. Blaxter, D. McK. Bird, and J. P. McCarter. 2005a. Comparative genomics in nematodes. Trends in Genetics, in press.

Mitreva, M. D., A. A. Elling, M. Dante, A. P. Kloek, A. Kalyanaraman, S. Aluru, S. W. Clifton, D. McK. Bird, T. J. Baum, and J. P. McCarter. 2004. A survey of SL1-spliced transcripts from the root-lesion nematode *Pratylenchus penetrans*. Molecular Genetics and Genomics 272:138–148.

Mitreva, M., J. P. McCarter, P. Arasu, J. Hawdon, J. Martin, M. Dante, T. Wylie, J. Xu, J. E. Stajich, V. Kapulkin, S. W. Clifton, R. H. Waterston, and R. Wilson, 2005b. Investigating hookworm genomes by comparative analysis of two *Ancylostoma* species. BMC Genomics 6:58.

Moens, T., and M. Vincx, 1998. On the cultivation of free-living marine and estuarine nematodes. Helgoländer Meeresunters 52:115–139.

Mungall, D. 2004. rNAPc2—Nuvelo. Current Opinion in Investigational Drugs 5:327–333.

Ott, J. A., M. Bright, and S. Bulgheresi. 2004a. Marine microbial thiotrophic ectosymbioses. Symbiosis 36:103–126.

Ott, J. A., M. Bright, and S. Bulgheresi. 2004b. Symbioses between marine nematodes and sulfur-oxidizing chemoautotrophic bacteria. Oceanography and Marine Biology: An Annual Review 42:95–118.

Parkinson, J., M. Mitreva, C. Whitton, M. Thomson, J. Daub, J. Martin, R. Schmid, N. Hall, B. Barrell, R. H. Waterston, J. P. McCarter, and M. L. Blaxter. 2004. A transcriptomic analysis of the phylum Nematoda. Nature Genetics 36:1259–1267.

Philippe, H., N. Lartillot, and H. Brinkmann. 2005. Multigene analyses of bilaterian animals corroborate the monophyly of ecdysozoa, lophotrochozoa, and protostomia. Molecular Biology and Evolution 22:1246–1253.

Salzberg, S. L., A. L. Delcher, S. Kasif, and O. White. 1998. Microbial gene identification using interpolated Markov models. Nucleic Acids Research 26:544–548.

Scholl, E. H., J. L. Thorne, J. P. McCarter, and D. McK Bird. 2003. Horizontally transferred genes in plant-parasitic nematodes: A high-throughput genomic approach. Genome Biology 4:R39.1–R39.12.

Stanssens, P., P. W. Bergum, Y. Gansemans, L. Jespers, Y. Laroche, S. Huang, S. Maki, J. Messens, M. Lauwereys, M. Cappello, P. J. Hotez, I. Lasters, and G. P. Vlasuk. 1996. Anticoagulant repertoire of the hookworm *Ancylostoma caninum*. Proceedings of the National Academy of Sciences 93:2149–2154.

Stein, L. D., Z. Bao, D. Blasiar, T. Blumenthal, M. R. Brent, N. Chen, A. Chinwalla, L. Clarke, C. Clee, A. Coghlan, A. Coulson, P. D'Eustachio, D. H. Fitch, L. A. Fulton, R. E. Fulton, S. Griffiths-Jones, T. W. Harris, L. W. Hillier, R. Kamath, P. E. Kuwabara, E. R. Mardis, M. A. Marra, T. L. Miner, P. Minx, J. C. Mullikin, R. W. Plumb, J. Rogers, J. E. Schein, M. Sohrmann, J. Spieth, J. E. Stajich, C. Wei, D. Willey, R. K. Wilson, R. Durbin, and R. H. Waterston. 2003. The genome sequence of *Caenorhabditis briggsae:* A platform for comparative genomics. PLoS Biology 1:E45.

Stein, L. D., C. Mungall, S. Shu, M. Caudy, M. Mangone, A. Day, E. Nickerson, J. E. Stajich, T. W. Harris, A. Arva, and S. Lewis. 2002. The generic genome browser: A building block for a model organism system database. Genome Research 12:1599–1610.

Stein, L. D., P. Sternberg, R. Durbin, J. Thierry-Mieg, and J. Spieth. 2001. WormBase: Network access to the genome and biology of *Caenorhabditis elegans*. Nucleic Acids Research 29:82–86.

Summers, R. W., D. E. Elliott, J. F. Urban, R. Thompson, and J. V. Weinstock. 2005. *Trichuris suis* therapy in Crohn's disease. Gut 54:6–8.

Tetteh, K. K., A. Loukas, C. Tripp, and R. M. Maizels. 1999. Identification of abundantly expressed novel and conserved genes from the infective larval stage of *Toxocara canis* by an expressed sequence tag strategy. Infection and Immunity 67:4771–4779.

The *C. elegans* Sequencing Consortium. 1998. Genome sequence of the nematode *C. elegans:* A platform for investigating biology. Science 282:2012–2018.

Whitton, C., J. Daub, M. Quail, N. Hall, J. Foster, J. Ware, M. Ganatra, B. Slatko, B. Barrell, and M. Blaxter. 2004. A genome sequence survey of the filarial nematode *Brugia malayi:* Repeats, gene discovery, and comparative genomics. Molecular and Biochemical Parasitology 137:215–227.